UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

# AI-Extended Moral Agency?

**Telakivi, Pii; Kokkonen, Tomi; Hakli, Raul; Mäkelä, Pekka**

**2026-01-02**
Routledge
http://hdl.handle.net/10138/595589
Telakivi, P, Kokkonen, T, Hakli, R & Mäkelä, P 2026, 'AI-Extended Moral Agency?', Social Epistemology : a journal of knowledge, culture and policy, vol. 40, no. 1, pp. 116-128. https://doi.org/10.1080/02691728.2025.2472759

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

Routledge
Taylor & Francis Group

Social Epistemology

A Journal of Knowledge, Culture and Policy

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tsep20

# AI-Extended Moral Agency?

Pii Telakivi, Tomi Kokkonen, Raul Hakli & Pekka Mäkelä

**To cite this article:** Pii Telakivi, Tomi Kokkonen, Raul Hakli & Pekka Mäkelä (16 Apr 2025): AI-Extended Moral Agency?, Social Epistemology, DOI: 10.1080/02691728.2025.2472759

**To link to this article:** https://doi.org/10.1080/02691728.2025.2472759

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 16 Apr 2025.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tsep20

SOCIAL EPISTEMOLOGY
https://doi.org/10.1080/02691728.2025.2472759

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

# AI-Extended Moral Agency?

Pii Telakivi[a], Tomi Kokkonen[a], Raul Hakli[a] and Pekka Mäkelä[a,b]

[a]Department of Practical Philosophy, University of Helsinki, Helsinki, Finland; [b]Helsinki Institute for Social Sciences and Humanities, University of Helsinki, Helsinki, Finland

**ABSTRACT**

In this paper, we ask how 'cognitive extenders', based on AI technology, affect their users' status as moral agents and the moral evaluation of their actions. We study how 'AI-extenders' can either enhance or diminish their users' moral agency. On the one hand, they can broaden the scope of agential features and on the other hand, they can undermine the agent's autonomy and lead to decreased responsibility. Our focus is on moral agency and responsibility of the AI-extended human being as a hybrid, coupled system. Assuming standard conditions for responsible agency, we will look at specific cases where the extender would make a difference to the agent's moral status. The thought-experimental extenders we are dealing with are enabled by already existing technologies. The obvious motivation behind the exercise is that these devices might be useful for people suffering from psychiatric conditions complicating the expression of their moral agency. We analyze the moral status of AI-extended agents as coupled systems and argue that the functioning of an AI-extender can make a difference to an agent's fitness to be held morally responsible. This should be considered in the responsible design and development of AI-extenders.

**ARTICLE HISTORY**
Received 5 February 2025
Accepted 15 February 2025

**KEYWORDS**
Cognitive extension; AI-extenders; moral responsibility; moral agency

## 1. Introduction

Throughout history, humans have off-loaded cognitive tasks to external tools and devices, but the recent advancements in AI and machine learning technology have elevated this phenomenon to an entirely new level. So-called AI-extenders (Hernández-Orallo and Vold 2019; Vold and Hernández-Orallo 2022) – AI-devices that perform substantial cognitive tasks as a part of their users' cognitive processes – carry a lot of promise. For example, such devices might be beneficial for people suffering from psychiatric conditions that complicate the expression of their moral agency. At the same time, they pose deep philosophical questions related to the 'mind-technology problem' concerning how new technologies shape our agency, personal identity and social practices, and what sets our minds apart from the technology we use (see Clowes, Gärtner, and Hipólito 2021). In this paper, we explore one aspect of the problem by looking at how AI-extenders provide opportunities and complications affecting our moral agency.

We begin by introducing the idea of cognitive extension and AI-extenders. In order to study the relevance of AI-extenders to assessments of moral responsibility, we describe the folk psychology view of what constitutes action, agency, and moral agency. We then provide examples of how devices with different capacities could boost or diminish the defining properties of moral agency, and hence potentially affect those folk psychological assumptions. Our discussion employs thought

**CONTACT** Pii Telakivi pii.telakivi@helsinki.fi Department of Practical Philosophy, University of Helsinki, P.O. Box 24, Helsinki 00014, Finland

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

experiments involving AI-extenders to highlight the potential impact of such devices under varying conditions. The thought experiments in the focus of this short paper instantiate and offer fresh insights into various aspects of the broad and multifaceted mind-technology problem (Clowes, Gärtner, and Hipólito 2021). Moreover, we hope this discussion will provide tools to address the question of what is responsible design and use of devices like AI-extenders. Hence, this paper can also be seen as an exercise in the value-sensitive design approach (Friedman and Hendry 2019) – we aim to bring into light potential risk types related to AI-extenders that should be considered in designing real-world AI-extended technology.

## 2. AI-Extended Cognition

In recent years, philosophers have become increasingly interested in devices that operate with AI and can become integrated or incorporated as part of the user (see e.g. Andrada, Clowes, and Smart 2022; Clowes 2015, 2019; Clowes, Smart, and Heersmink 2024; Naeem and Hauser 2024; Wheeler 2019). Particular attention has been drawn to ethical questions, such as mental privacy, the risk of manipulation, the risk that arises from a high level of phenomenological transparency without understanding how the technology works, and the potential undermining of mental autonomy. We have chosen to use AI-extenders (Hernández-Orallo and Vold 2019; Vold and Hernández-Orallo 2022) as our central conceptualization, because it is particularly useful when examining how questions of moral agency and responsibility change as AI technologies become integrated into our minds. This concept is a direct continuation of Clark and Chalmers' (1998) hypothesis of Extended Cognition that states that external technologies can become constitutive parts of cognitive functions like memories and beliefs.[1]

In contrast to standard discussions on extended cognition that focus on cases like using pen and paper to boost one's memory or using a tourist map for navigation in a new city, the tools under scrutiny are machine learning systems that realize parts of our cognitive functions. Hernández-Orallo and Vold (2019) define AI-extenders as cognitive extenders that use AI technology, so that a new cognitive capability is acquired. Unlike autonomous AI systems (like a delivery robot), they are non-autonomous parts of a hybrid system that is autonomous as a whole. They are devices, apps and wearables that can use auditory, visual or tactile recognition technology, and they can be used for various purposes, for example, to help people who suffer from learning disorders or mental disorders.

As an example of an AI-extender, consider a taxi driver using a navigation app powered by AI technology. She always has it on when she drives and doesn't question its instructions when everything seems to go well. The interface she is familiar with is the app, and not the street signs or other traditional navigation elements. Her navigation skills have evolved through using the app: she has become very competent at navigating the city – but were the app to be removed, her navigation skills would be negligible. The navigator might also make her a safer driver, since it can inform her of speed limits and routes to avoid. As a coupled system, the driver and the navigation app form a unit that we are interested in when talking about AI-extenders (for similar examples in the literature, see e.g. Andrada, Clowes, and Smart 2022; Clowes 2020).

New technologies have transformed not only navigation (Gillett and Heersmink 2019), but many other everyday tasks and cognitive functions, such as memory (Clowes 2017; Heersmink and Adam Carter 2020), and also larger phenomena, such as our social relationships, work practices, leisure activities, fitness routines, and healthcare. Accordingly, ethical issues like moral responsibility have become more complex and pressing. For example, privacy concerns are pronounced, because cognitive aids based on machine learning technology can potentially be controlled by third parties that collect data about their users, leaving the agent unable to control who has access to the contents of her mental life (there is a vast literature about privacy concerns that arise from the extended mind, see e.g. Carter 2021; Clowes, Smart, and Heersmink 2024). Further, since AI-extenders have more potential to develop control over their users, there is a risk of endangering

autonomy and engendering cognitive atrophy (Hernández-Orallo and Vold 2019). Nevertheless, AI-extenders might also empower individuals who lack certain abilities (e.g. recognizing emotions via facial expressions) by compensating for or augmenting them.

To illustrate the dilemma we want to tackle in this paper, consider Ilari, a young athlete diagnosed with ADHD. He practices his sport alongside a group of teammates. Rather often, Ilari is taken by an 'anxiety seizure'. In his own words, his head is 'full of ants'. This uncomfortable predicament may drive Ilari to erratic behavior, bullying and disturbing other teammates to relieve his anxiety. However, Ilari is on the whole a well-intentioned young man, and he usually tries to resist the urge 'to go mad'. Sometimes he succeeds, with an inhuman effort. His understanding coach praises him when he succeeds in staying calm, and attributes the bad behavior to the condition rather than to Ilari himself, and thus tends to react without blame.

One day, Ilari is given a new device in the form of, say, an AI-based smartwatch connected to other wearable smart technology that assists with anxiety management by detecting the physiological features and providing warnings like a buzz connected to a verbal reminder to take particular calming actions. He starts using it regularly when training and gradually becomes accustomed to it. The teammates report that Ilari is like a different person – for the better! He understands quicker that an anxiety attack is on the way and takes the recommended course of action to reduce negative behavior.[2]

Does 'new Ilari' deserve more or less credit for not beating up his teammates (compared to earlier rare successes), because controlling himself doesn't require as much effort? Or should his good behavior be valued as before, since his intention hasn't changed, it has only become easier to accomplish? In other words, how does the device affect his praiseworthiness or blameworthiness for anxiety management? How should we understand the mind-technology conundrum – including intentional action, agency, and moral responsibility – with technology-amplified agents like Ilari? To answer this question, we will now delve into the domains of action, agency, and moral responsibility, and pose a further question: where should cognitive extenders be positioned within them?

## 3. Locating AI-Extenders in the Moral Responsibility Debate

How should we understand moral agency and responsibility? In a broad sense, actions can be understood as manifestations of agency, and this creates the basis for moral responsibility. If an agent intentionally does something she knows to be morally wrong, is not coerced to perform that action, and there is no external pressure of any kind present, then it is justified to ascribe moral responsibility for that action to her. Responsibility requires the powers of reflective self-control: the power to grasp and apply moral reasons and the power to control or regulate her behavior by the light of such reasons. Furthermore, the expression of the agency should not be 'disturbed' by *exempting conditions* (such as serious mental disorders, extreme immaturity, psychopathy) nor *excusing conditions* (such as inadvertence, mistake, coercion).

The elements of intentional action and agency are described under a different conceptual framework than the causal–cognitive processes that ground them. These levels of description should not be conflated, but the details of cognition are relevant to how we analyze agency. Cognitive processes are environmentally situated, and external elements such as technological aids we use play a substantial role. There is an ongoing debate about whether we should understand cognition as *embedded* in a network of causal influences or as partly *constituted* by external elements, that is, as *extended* cognition (Clark and Chalmers 1998; Rupert 2009; Telakivi 2023). The constitution of agency, however, is a separate issue although not independent from both ontological and explanatory questions about cognition. Regardless of whether we take cognitive processes, foundational to agency, as partly constituted by external tools or merely causally influenced by them, the agency is still dependent on these external tools. But it is a separate matter to determine whether the external tools are a constitutive part of the agent. This has to be analyzed from an action-theoretical

perspective, sensitive to the very aims of this conceptual framework in judging whether the agent is in control and not, for example, manipulated in their behavior.

Our interest is in action and ethics, and what matters from this perspective is that AI-extenders may affect different parts of the chain in the intentional action description (intention formation, reasons-sensitivity, intention execution) and, consequently, the elements that constitute moral agency. AI-extenders may influence whether an exempting or excusing condition applies or no longer applies. The identification of the locus of the impact in the action chain is relevant to the moral responsibility appraisal. AI-extenders can strengthen or weaken the moral status of their users, and even change their status as morally responsible agents.

There is a lively debate concerning the role of moral considerations in deciding the ontological question of whether external tools should be understood as constituting or causally affecting the cognitive agent (see e.g. Cassinadri 2022; Clowes, Smart, and Heersmink 2024; Farina and Lavazza 2022). As illustrated in the example in Section 2, the moral evaluation of an action performed by a morally responsible agent like Ilari who is wearing a smartwatch may depend on whether we count the smartwatch as a constitutive part of the agent or not. However, in the following, our focus is on cases in which not only the evaluation of an agent's action but the agent's *status as a moral agent* may depend on the presence of an AI-extender and how that might affect whether we should identify the extender as a constitutive part of the agent or not.

Traditionally, the type of moral agency that interests philosophers has been a morally responsible agency. To be a morally responsible agent is to be an appropriate subject of the normative demands and expectations through the practices of holding agents accountable of the moral community (Darwall 2009). Here, we adopt what we take to be the standard view of responsible agency, involving at a minimum, a certain degree and quality of control (Fischer and Ravizza 1998; McKenna 2012; Wallace 1994). Whatever one's view of these capacities and abilities, though, they are taken to be those possessed by mature, neurotypical adults (Strawson 1962). Thus, even among those who share the widely accepted view that capacities required for moral agency come in degrees, morally responsible agency is generally thought to arise only at a certain threshold of maturity and control (e.g. a sufficient sensitivity to reasons, cf. Fischer and Ravizza 1998).

Recently, there has been plenty of discussion over the question whether AI agents could be morally responsible (Floridi and Sanders 2004; Himma 2009; Johnson 2006; Sullins 2006; see, e.g. Hakli and Mäkelä 2019) and there has also been discussion on whether humans and AI agents could share moral responsibility for consequences that ensue from the interaction between a human and an AI system (Hakli and Mäkelä 2019; Neuhäuser 2015). However, the question of whether extended agents composed of individual human beings who employ AI-extenders can be held morally responsible remains unanswered. As noted by Hernández-Orallo and Vold (2019), when coupled, an agent, say A, and an AI-extender, say E, form an extended agent A[E] that is different from a collective agent A+E. The action of such extended agents is not collaboration or shared agency, since the extender itself is not autonomous. These new technologies have opened up this whole-new dusky area between tool use, collaboration and hybrid agency that the mind-technology problem tries to navigate. In the next section, we will analyze situations in which the use of an AI-extender may affect one's status as a moral agent with respect to a certain domain or context of action.

## 4. AI-Extenders Alter the Status of Moral Agency

We will now consider cases in which the presence of an AI-extender makes a difference to an agent's status as a moral agent capable of bearing moral responsibility for her actions, possibly concerning only some particular domain of action. For the purposes of this paper, we adopt a rather typical set of conditions for moral agency which are slightly modified from those presented by Himma (2009). Accordingly, for all agents X, X is a moral agent if and only if X is

(1) capable of making free choices,
(2) capable of deliberating about what one ought to do in light of moral reasons, and
(3) motivated and able to act on the basis of such deliberation.

In the following, we will ask what impact the use of an AI-extender has on the agent's capacities and consequently on the fulfillment of conditions 1–3. If the extended agent simply inherits the capacities of the bare agent, then the extender doesn't affect her moral status. Interesting cases would be those where adding or removing an extender makes a difference to whether the agent is a moral agent or not.

The extenders we use as examples are plausible but hypothetical devices: the necessary technical preconditions for many of them exist already, but as such most of them are not yet in use. The first three deal with psychiatric conditions (based on the diagnostic manual DSM-5, see APA 2022). This relates to real-life needs, because these technologies might have a considerable effect on certain kinds of deficiencies (and hence be useful for an individual with said deficiencies). However, these devices could also be useful to anyone who wishes to modify her set of morally relevant capacities. We certainly don't claim that these devices would fit only one diagnostic category – on the contrary, many diagnostic categories exist on a spectrum (and depend on classificatory and cultural factors), and in a similar manner as other treatments in psychiatry, these extenders might benefit one while not someone else.

The conditions listed above are categorical criteria for whether a person is a moral agent or not, but the capacities themselves are not either/or properties. They may come in degrees in several dimensions. For instance, a capacity can have different *strengths* depending on how reliably or robustly it manifests in a person's action or thinking in general. Additionally, the capacity can have different *scopes* in the sense that the capacity may exist (or function properly) in a wider or narrower range of contexts and external conditions. The practical relevance of extenders being able to affect moral agency comes from the fact that the relevant capacities can be strengthened or weakened, or their scope can be affected; the demonstration of this ability comes from the analysis of what kind of role they could have in enabling or disabling moral agency of a person. We will discuss examples of both kinds. Furthermore, the psychological capacities satisfying the conditions may be constituent parts of more than one condition, but we will consider each condition separately.

### *4.1. AI-Amplified Moral Agency*

To begin with, we will introduce three different ways an extender could potentially affect its user's status as a moral agent by enhancing the agent's capacities to the extent that the above conditions 1–3 become satisfied.

First, let us introduce a person called Mari who has symptoms of borderline personality disorder (BPD), and has difficulties in making morally relevant decisions on her own. BPD patients often have deficiencies in regulating emotions and moods, and this lack of self-regulation and impulse control can undermine their decision-making abilities. For the sake of an argument, let us imagine that Mari is suffering from such a severe case of BPD that her capacity for making free choices is impaired to the degree that the first condition for moral agency doesn't hold for her.

When Mari is coupled with an AI-extender, let's say wearable smart technology designed to enhance metacognitive skills by improving impulse control, the situation changes. The 'impulse control shirt' can increase her concentration and decision-making capacities to the extent that she is now a competent moral decision-maker regarding a new domain (emotional self-regulation). Hence, the first condition that appealed to the capability of making free choices didn't hold with her to start with, but it does after coupling with the AI-extender. We can conclude that the extender made a difference to whether she can be considered a moral agent.

As a second character, consider Riku who has difficulties in interpreting other people's behavior and is incapable of reading and recognizing their emotions. This could be a real

situation for a person with an autism spectrum condition. The inability to read emotions in other people's behavior makes Riku insensitive to certain morally relevant aspects of social situations. For example, Riku often fails to notice that he has upset someone, and that makes him incapable of deliberating what he should do in those situations. However, he would very much want to appreciate other people's emotions, if only he could recognize them. Due to his 'blind spot' – a domain-specific shortcoming – he doesn't globally fulfill the second requirement of moral agency.

Riku might benefit from an extender, say, in the form of a smart helmet that can detect emotion features in facial gestures, movements and tone of voice. It would enable Riku to react in accordance with other people's emotions and make his life much easier. When Riku starts to wear the helmet habitually, he learns to access the information provided by the extender in his deliberations and take appropriate actions based on it. After coupling, he is considered morally responsible for his behavior in a new class of social situations. In other words, Riku satisfies the second condition of moral agency more globally with an AI-extender than without, thus the extender has strengthened Riku's moral agency.[3]

Third, consider Lisa who can reason about what she should do and choose a suitable action plan accordingly, but cannot find sufficient motivation to initiate the action, because she suffers from depression. She doesn't have problems understanding what she ought to do and make plans based on that, but she is listless and lacks motivation to do anything. She satisfies the first two conditions, but the third condition (motivation and ability to act based on moral reasons) fails without an extender.

A suitable AI-extender for Lisa could be a predictive medical tool ('a motivation pump') that helps monitor its user's moods and recommends the best timing for certain actions. It monitors when the user is at her strongest mindset, and in contrast, when morally challenging situations are better to be avoided if possible. It recognizes when Lisa is in critical situations where she undergoes moral deliberation, decides on the appropriate action, and *tries* to initiate action (but fails without the device). Like an insulin pump increases insulin when needed, a motivation pump increases self-confidence when needed. The pump injects critical chemicals to boost her levels of endorphins and dopamine that increase her subjective feeling of well-being thereby giving a positive reward and reinforcing her motivation to act. Hence, it seems that the AI-extender might enhance the capacity to be motivated by moral reasons, and condition three would hold with, but not without an AI-extender. However, strengthening the motivation is probably more problematic and difficult compared to the first two cases. Motivation is undoubtedly harder to compensate for than metacognitive skills or recognition of emotions.

All these three examples display cases of successful compensation. The use of an AI-extender affected their status as morally responsible agents with respect to a certain domain. The extender made a difference to whether the conditions of moral agency were satisfied: Mari became a moral agent regarding the capability of making free choices; Riku regarding the capability to deliberate based on moral reasons; and Lisa regarding the capability to be motivated to act based on moral reasoning. They were not moral agents to start with in these contexts of action, but when the extender was added an exempting condition got eliminated, and they became moral agents also in these new domains.

### *4.2. AI-Diminished Moral Agency*

Let's turn the tables and consider examples from the opposite perspective where the AI-extender would undermine its user's moral agency. The following examples introduce three characters who are moral agents to start with, that is, they fulfill all three conditions of moral agency. But when they are coupled with an extender, their moral agency will diminish in certain domains. We have chosen these examples to show that besides positive effects, AI-extenders might also have a negative impact on their user's moral agency. The initial aim to start using these technologies is to augment one's

capacities, but it can lead to unwanted outcomes. Whereas in the previous cases the extender withdrew an exempting condition, in the following examples, the extender creates one. So, let us introduce our last three protagonists.

As our fourth AI-extender user, may we present Reko, who has felt, recently, that he could do with some help in dealing with his daily cognitive burden, such as keeping up with his busy work schedule, taking his kids to their hobbies and keeping up communication with his old friends. He decides to give a try to a 'cognitive assistant' based on AI-technology. It is an app on his smartphone, connected to wearables that constantly prompt him to act a certain way in order to boost efficiency in planning his daily cognitive tasks. It can plan, allocate and schedule tasks in ways that he wouldn't have figured out without it, it can suggest responses and ready-made templates for (at least written) communication, and if his daily schedule seems to be too full to be able to cope with, sometimes the extender might advise to give up some tasks altogether.

First, all goes well, but after some time, Reko starts to feel that he is no longer in charge of his decisions. The cognitive assistant takes care of his daily decisions to the point where his choices result from the device, not from his free choice. He is manipulated by the device, and his own will has become undermined. The extender has an effect that Reko is no longer in a position where he could be said to have full control of his choices, so we can conclude that it undermined his status as a moral agent.

Fifth, consider Jari, who has consumed a lot of self-help literature convincing him that he could become a better version of himself. He buys a device that is advertised as a memory enhancer, which improves memory, provides a boost both in terms of the quantity of items one can recall, and also in terms of the detail with which one is able to bring back long-lost incidents from his past. Tempting, indeed, and first he enjoys the door that has opened to all those memories.

But sometimes the technology misfires and feeds him lively images of events that never took place and associates incidents and people from his life in non-factual ways – in a similar way as large language models like ChatGPT sometimes hallucinate 'facts'. Eventually, the extender distorts Jari's picture of reality, the details of the shared histories of the people around him, and his own narrative identity. Contaminated with these false beliefs, Jari starts to base his moral pondering on reasons that would never have directed his actions before. Consequently, he is not capable of morally adequate deliberation, and his status as a moral agent is undermined.

Finally, Sari, who has always been a well-behaved and dutiful person who wants to do morally right things. However, after some difficult life experiences, she feels that she wouldn't always want to keep up the effort to be a good person. She decides to buy a 'moral enhancer' in the hope that she can outsource morally demanding decisions to the device. First, everything goes well, she feels less burdened but with the same morally praiseworthy outcomes. However, after using the enhancer for a while, she starts to lack motivation to act based on its prompts, because she feels they are not 'genuinely hers'. She feels that the extender has turned her into a 'right-doer', but at the same time, her own motivation to do the right things has been undermined. The extender has affected the third condition of moral agency, since she no longer has the same motivation to act based on the moral reasoning as before using the enhancer.

These 'negative' cases show that AI-extenders can also undermine moral agency. Reko has lost the capability to make free choices, Jari has lost the capability to pursue moral reasoning based on relevant information, and Sari has lost her motivation to act morally. More generally, the intrusion into their moral capacities can lead to a feeling of alienation from oneself as an autonomous moral agent. How should we approach moral appraisal in these cases? It seems quite unfair to require something that is not even within one's grasp. The answer depends on whether one decides to opt-in with the augmenting AI-device in the first place and is aware of the risks involved. Also, as the use of these technologies becomes more widespread in society, it becomes increasingly difficult to avoid using them, as so many essential functions begin to rely on them (just as it is very challenging to navigate society today without a smartphone).

Of course, Reko, Jari, and Sari would probably get their lost capacities back if they quit using their extenders. Indeed, it is important that there is a possibility to opt out of this kind of highly pervasive technology. However, there are potential obstacles on the way of withdrawing from cognitive extenders. First of all, they can be very addictive: it is hard to return to one's old habit that requires more willpower. Moreover, sometimes extender usage might lead to cognitive atrophy. Reko may find it difficult to regain control and get back to making his own choices, and Jari might permanently stick with distorted views about his past. The displacement of heavily relied-on technology might lead to a significant deterioration in the functions that were outsourced to it. We might soon have a generation of students who have never written an essay without ChatGPT. Also, recall our taxi driver presented earlier. If she never navigates without an app, her navigation skills will degenerate. The phenomenon of cognitive atrophy is a real worry with our current lifestyle. This is how new technologies have always worked, of course, but some of the capacities we have lost might not be necessary for our well-being whereas our capacity to make moral decisions might be too important to lose.

## 5. Reconsidering the Attribution of Moral Responsibility

Now, let's analyze how to attribute moral responsibility when a person is coupled with an AI-extender. As our cases showed, there's no one-size-fits-all answer, but we need to approach this question case by case. In the positive cases, the extenders served as cognitive tools, facilitating Mari's, Riku's and Lisa's control over their own actions and moral judgements. They are responsible agents because their extenders enabled the realization of the conditions of moral agency for them. In the negative cases, the extenders distorted Reko's, Jari's and Sari's autonomous decision-making, capacity for moral reasoning, or the ability to be motivated to act according to their deliberations. Coupling with their extenders thwarted the fulfillment of the conditions of moral agency, and hence they cannot be judged as before.

If one feels controlled by the technology, rather than using it to achieve intentionally chosen goals, it alienates one from their autonomous agency. In the first three examples, Mari, Riku, and Lisa used extenders successfully. However, even in these successful cases, it is questionable whether they had control (and a feeling of control) over it. When a painter uses a brush or a carpenter uses a hammer, they are in control of the tool, and not the other way around. However, it seems that we are facing a different situation when the tool is powered by AI or machine learning technology, or some other type of smart technology that has high adaptability. They are tools that don't resemble what we are used to thinking tools are. They come with much higher degrees of adaptability and individualization, leading to outcomes that are far more unpredictable than those with traditional tools. They are *transparent-in-use* (Andrada, Clowes, and Smart 2022), but we often don't know how they work or, indeed, how they affect us as human agents. They fulfill roles that resemble social functions – they indeed have many skills that only a few years ago were preserved for humans alone. They can recognize emotions; produce language, pictures, and music; keep company and engage in a conversation. Introducing AI tools into the cognitive system may cause the human in the loop to lose the 'final word'.

Our examples highlighted the possibility that both internal and external features may be significant in evaluating and determining whether the target agent satisfies the conditions of a moral agent. In all these cases, the use and operation of the device are issues of moral deliberation in themselves. If something goes wrong, moral responsibility may lie with the designer; on the other hand, in some circumstances, the responsibility may lie with the user. In any case, the responsible design and implementation of these technologies call attention to highly important ethical questions.

There is an asymmetry between the positive cases, in which the technology enables moral agency, and the negative cases, in which it seems to undermine it. In the positive cases, we have no problem in saying that the person-technology system is a moral agent in a way the person without technology would not be, while in the latter case, it seems more intuitive to say that the person is a moral agent whose ability to manifest agency in action is prevented by the technology. In

other words, we have different intuitions about whether to include the extender into the system that underlies the agency or is outside it. Yet both kinds are examples of systems consisting of a person and an extender of a particular cognitive capacity. Is there any reason to treat the two sets of cases differently?

As discussed in Section 3, the constitution of a cognitive system and the constitution of agency are distinct issues, even though being an agent depends on having the appropriate cognitive capacities. This means that even if we describe both systems symmetrically at the *cognitive level of description*, we may have reasons to describe them asymmetrically at the *agentive level of description* (see Kokkonen 2021 for discussion on the levels of mental description). Much of the debate about whether an external element in a cognitive process is simply a tool for embedded cognition or a constitutive component of extended cognition has been done in terms of *mechanistic explanation* (see e.g. Kaplan 2012; Krickel 2023).[4] According to the mechanistic approach, a cognitive process has parts that constitute it, but it may also be in causal interaction with something that is not its constitutive part but is nevertheless (causally) enabling the process. The mechanist approach takes an explanatory perspective – it promises to offer a scientific explanation rather than a metaphysical enquiry into the ontological nature of the phenomenon in question. However, it has become clear that mechanistic analysis alone cannot solve the causal–constitution debate in the extended cognition framework, as many of its critics have shown. In many cases, it is possible to describe the role an extender plays in a cognitive process in both causal and constitutive terms. However, as many proponents of extended cognition have suggested (see e.g. Hurley 2010), a better way to demarcate cognitive and agential boundaries can be achieved with a case-by-case method.

Moral agency is built on the set of capacities as outlined in Section 4. An entity that has these capacities is regarded as a moral agent. We can reverse this: if a moral agent is to be identified, those capacities that make one, belong to the agent. This provides different criteria for different cases to identify the elements that are constitutive parts of the cognitive processes of the agent and which are causal factors. This would create an asymmetry between the two sets of cases: in the positive cases, the extenders participate in creating moral agency, while in the negative cases, there is a moral agent suppressed by technology. Wheeler (2019) has suggested that we should take a step back from extended cognition to embedded cognition in certain potentially risky cases of AI-extended cognition. Should we take a similar approach with our 'negative cases'? But as Clowes, Smart, and Heersmink (2024) note, the biggest open questions of responsibility, autonomy, privacy, and mental manipulation do not change at all, regardless of whether we classify AI-extenders as causal or constitutive elements. They continue that the proponent of the extended view would say that the embedded view overlooks important ethical issues related to individual autonomy and dignity, if the technologies are not considered as proper parts of their users. Conversely, the proponent of the embedded view would say that the extended view exaggerates ethical claims about devices that, after all, are material artifacts, equating them with issues concerning the autonomy of persons (Clowes, Smart, and Heersmink 2024). Instead of seeing the agent and her AI device in a vacuum, we should take into account the context she is in – the whole distributed cognitive network.

AI-extenders indeed have the power to affect our agency, including moral agency, in both positive and negative ways. For example, a face recognition app can be very useful for someone who suffers from prosopagnosia and lives in the EU, but very dangerous for someone living in an authoritarian state. AI-extenders ethical classification always depends on what exactly their functions are, who is using them, what kind of other technical, cognitive, and moral skills the users possess, and in what kind of society and social situation they are in. This entails grave moral responsibility in the design process and highlights the importance of responsible design accounting for the range of contexts in which the technology is used, not just what it generally speaking does.

## 6. Conclusions

We have argued that AI-extenders can affect the status of their users' moral agency and hence have an impact on issues of responsibility. They may both enable and disable moral agency, depending on the type of extender and the context. We hope to have paved the way for further research on the responsible design and use of AI-extenders. In terms of the enabling part, AI-extenders can both compensate and enhance cognitive abilities and lead to alterations in the agent's moral status. One could argue that the former (compensating or substituting an ability as a form of treatment) is morally more acceptable than the latter (enhancing or augmenting the agent by creating new abilities). However, it is impossible to draw a neat line between the two, given the lack of a definable range of normal variation in human cognitive capacities. The distinction between compensating a deficiency and enhancing a capacity to get extraordinary benefits becomes an issue of justice, not normality. Instead, we suggest that we should think about AI-extender technology as a normative and practical issue. Where do we need AI-extenders and why? In which cases are they especially harmful and threaten to undermine mental autonomy and moral agency? And what would be the best way to educate the general public about their risks and benefits? Whether we want it or not, they are here to stay and will continue to challenge our previously held concept of human mind and moral agency.

Finally, the main benefit that arises from the AI-extended scope of moral agency is not just about social justice. It is about enabling people to become fully capable agents in the moral dimension as well as ensuring they are capable of meeting the expectations of their society. Ilari, the young athlete who suffers from ADHD, would both feel out of control of himself and be an outcast without the device. Coupled with his extender, he can express his full moral agency, without the burden of the extra task of overriding an obstacle that would make common and easy ethical duties supererogatory for him.

AI-extenders can affect an agent's decision-making and decisions about their technical functions, design, as well as who can use them and where, are not morally neutral. They may both enable an individual to be a responsible agent and nudge the content of the choices at the same time. While this technology could extend the scope of moral agency in one dimension, it could diminish autonomy in another, and lead to growing dependency. There are also worries about biases and abuse, as always with AI, and they are very intimate in this case. Developing psycho-technological hybrids holds great potential but also comes with significant risks. The design of these technologies requires careful oversight of the ethical consequences for the user and the social environment, case by case, application by application, customized for each user. We hope that we have provided some general philosophical insights for this value-sensitive design process.

## Notes

1. There has been a vast debate in the literature over the last 20 years whether (properly integrated) tools should indeed be classified as constitutive parts of cognition (this view is referred to as *extended* cognition) or merely as causal factors (this view is referred to as *embedded* cognition). We will return to this question in Sections 3 and 5.
2. Assistive smart technologies for children with ADHD have developed a lot in recent years, and there are applications that function much like Ilari's smartwatch. McHugh et al. (2010) designed a smart wristband connected to a heart rate belt that can detect upcoming emotional outbursts and advise the child on self-calming routines. For a review of other smart technologies in use for people with ADHD, see Sonne et al. (2016).
3. This kind of technology could also be used to detect one's *own* emotions, and that might also increase deliberation over morally relevant issues. Perhaps it could also help neurotypical individuals to recognize emotions in others in online environments. There are several kinds of technologies available for this purpose, e.g. a smartphone app that detects one's moods based on their typing patterns (Ghosh et al. 2019).
4. It is even likely that the whole 'constitution-turn' in the extended cognition literature took place precisely because of the success of the mechanistic view in philosophy of science (see Telakivi 2023). However, there have been other ways to explain the constitutive relation, most notably Michael Kirchhoff's account on *diachronic constitution* (see e.g. Kirchhoff 2015).

## Acknowledgments

We thank the anonymous reviewers for their comments that helped us to improve the quality of this article. We also want to thank Dane Leigh Gogoshin who was initially involved in our discussions where we began developing ideas for this article. This work was supported by the Kone Foundation under Grant number 201906341; NordForsk under Grant number 105081; and the Strategic Research Council (SRC) established within the Research Council of Finland under Grant number 353400.

## Disclosure Statement

No potential conflict of interest was reported by the author(s).

## Funding

The work was supported by the Koneen Säätiö [201906341]; NordForsk [105081]; Strategic Research Council [353400].

## Notes on contributors

***Pii Telakivi*** is a post-doctoral researcher in RADAR: Robophilosophy, AI Ethics, and Datafication Research at the University of Helsinki, and in Temporality in Predictive Processing at the University of Turku. Her research focuses on extended, embodied cognition, exploring the intersections between philosophy of mind, technology, AI, and psychiatry. Her monograph "Extending the Extended Mind: From Cognition to Consciousness" was published by Palgrave Macmillan in 2023. Her recent academic appointments include a Fulbright Scholarship at UC Berkeley (2022–2023) and visiting scholarships at Macquarie University in Sydney (March 2024) and at the University of Amsterdam (2025–2027).

***Tomi Kokkonen*** works as a University Lecturer in theoretical philosophy at the University of Helsinki. His research interests are in philosophy of science (especially in the issues emerging in the intersections between biology, psychology and social sciences), philosophy of technology (especially metaphysical, conceptual and normative issues related to technology as a part of social practices, as well as the possibility and nature of artificial mental phenomena), philosophy of mind, and philosophy of sociality. His PhD thesis (2021) was on the evolutionary explanations of human sociality. His current research involves ethical and societal issues related to AI and robotics as well as the possibility of moral machines.

***Raul Hakli*** works as a university researcher in practical philosophy at the University of Helsinki. He did his PhD thesis in 2010 on the nature and logic of group beliefs. He has worked for several years on issues of collective intentionality, social ontology, and collective epistemology, and recently, his focus has been on philosophy of technology, in particular responsibility issues stemming from technologies like AI and robotics. He is co-leader of the research group RADAR together with Pekka Mäkelä, and he has lead research projects on the topic of responsible AI. He is also the Editor-in-Chief of Springer series Studies in the Philosophy of Sociality.

***Pekka Mäkelä*** is the vice-director and a research coordinator in the Helsinki Institute for Social Sciences and Humanities (HSSH) at the University of Helsinki. His research interests are in normative dimensions of collective and social action, e. g. collective responsibility and trust, social ontology, the philosophy of the social sciences, and philosophical problems of social robotics and human-robot interaction. He has been a visiting fellow at the Centre for Applied Philosophy and Public Ethics and ANU, Canberra, Australia, and he has also taught as an adjunct teacher at FSU, Florida, USA. His publications include "The collectivist approach to collective moral responsibility" (with Seumas Miller, Metaphilosophy, 2005), "Collective Agents and Moral Responsibility" (Journal of Social Philosophy, 2007), Trust: Analytic and Applied Perspectives (ed. With Cynthia Townley, VIBS, RoDoPi, 2013), "Group Agents and Their Responsibility (with Raimo Tuomela, Journal of Ethics, 2016), "A realist account of the ontology of impairment" (with Simo Vehmas, Journal of Medical Ethics, 2008), and "Moral Responsibility of Robots and Hybrid Agents" (with Raul Hakli, The Monist 2019). Presently, he is co-leader of the RADAR group together with Raul Hakli.

## ORCID

Pii Telakivi http://orcid.org/0000-0002-2094-8646
Raul Hakli http://orcid.org/0000-0001-8201-6409
Pekka Mäkelä http://orcid.org/0000-0002-0854-522X

## References


Andrada, G., R. W. Clowes, and P. Smart. 2022. "Varieties of Transparency: Exploring Agency within AI Systems." *AI & Society* 38 (4): 1321–1331. https://doi.org/10.1007/s00146-021-01326-6.

Carter, J Adam. 2021. "Varieties of (Extended) Thought Manipulation." *The Law and Ethics of Freedom of Thought, Volume 1: Neuroscience, Autonomy, and Individual Rights*. 291–309. https://doi.org/10.1007/978-3-030-84494-3_10.

Cassinadri, Guido. 2022. "Moral Reasons Not to Posit Extended Cognitive Systems: A Reply to Farina and Lavazza." *Philosophy & Technology* 35 (3): 64. https://doi.org/10.1007/s13347-022-00560-0.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58 (1): 7–19. https://doi.org/10.1093/analys/58.1.7 .

Clowes, Robert W. 2015. "Thinking in the Cloud: The Cognitive Incorporation of Cloud-Based Technology." *Philosophy & Technology* 28 (2): 261–296. https://doi.org/10.1007/s13347-014-0153-z.

Clowes, Robert W. 2017. "Extended Memory." In *The Routledge Handbook of Philosophy of Memory*, 243–254. Routledge.

Clowes, Robert W. 2019. "Immaterial Engagement: Human Agency within the Cognitive Ecology of the Internet." *Phenomenology and Cognitive Science* 18 (1): 259–279. https://doi.org/10.1007/s11097-018-9560-4.

Clowes, Robert W. 2020. "The Internet Extended Person: Exoself or Doppleganger?" *Limité Límite Revista Interdisciplinaria de Filosofía y Psicología* 15 (22): 1–23.

Clowes, Robert W., Klaus Gärtner, and Inês Hipólito. 2021. "The Mind Technology Problem and the Deep History of Mind Design." In *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts*, edited by Robert W. Clowes, Klaus Gärtner, and Inês Hipólito, 1–45. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-72644-7_1.

Clowes, Robert W., Paul Smart, and Richard Heersmink. 2024. "The Ethics of the Extended Mind: Mental Privacy, Manipulation and Agency." In *Neuro-Prosthethics*, edited by J. Heinrichs, B. Beck, and O. Friedrich, 13–35. Berlin: J. B. Metzler.

Darwall, Stephen. 2009. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, Massachusetts: Harvard University Press.

Farina, Mirko, and Andrea Lavazza. 2022. "Incorporation, Transparency and Cognitive Extension: Why the Distinction Between Embedded and Extended Might be More Important to Ethics than to Metaphysics." *Philosophy & Technology* 35 (1): 10. https://doi.org/10.1007/s13347-022-00508-4.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Floridi, Luciano, and J. W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14 (3): 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Friedman, Batya, and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001.

Ghosh, Surjya, Kaustubh Hiware, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. "Emotion Detection from Touch Interactions During Text Entry on Smartphones." *International Journal of Human-Computer Studies* 130 (October): 47–57. https://doi.org/10.1016/j.ijhcs.2019.04.005.

Gillett, Alexander James, and Richard Heersmink. 2019. "How Navigation Systems Transform Epistemic Virtues: Knowledge, Issues and Solutions." *Cognitive Systems Research* 56 (August): 36–49. https://doi.org/10.1016/j.cogsys.2019.03.004.

Hakli, Raul, and Pekka Mäkelä. 2019. "Moral Responsibility of Robots and Hybrid Agents." *The Monist* 102 (2): 259–275. https://doi.org/10.1093/monist/onz009.

Heersmink, Richard, and J. Adam Carter. 2020. "The Philosophy of Memory Technologies: Metaphysics, Knowledge, and Values." *Memory Studies* 13 (4): 416–433. https://doi.org/10.1177/1750698017703810.

Hernández-Orallo, José, and Karina Vold. 2019. "AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 507–513. AIES '19. New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3306618.3314238.

Himma, Kenneth Einar. 2009. "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent?" *Ethics and Information Technology* 11 (1): 19–29. https://doi.org/10.1007/s10676-008-9167-5.

Hurley, Susan. 2010. "The Varieties of Externalism." In *The Extended Mind*, edited by R. Menary, 101–153. Cambridge, MA: MIT Press.

Johnson, Deborah G. 2006. "Computer Systems: Moral Entities but Not Moral Agents." *Ethics and Information Technology* 8 (4): 195–204. https://doi.org/10.1007/s10676-006-9111-5.

Kaplan, David Michael. 2012. "How to Demarcate the Boundaries of Cognition." *Biology & Philosophy* 27 (4): 545–570. https://doi.org/10.1007/s10539-012-9308-4.

Kirchhoff, Michael D. 2015. "Cognitive Assembly: Towards a Diachronic Conception of Composition." *Phenomenology and the Cognitive Sciences* 14 (1): 33–53. https://doi.org/10.1007/s11097-013-9338-7.

Kokkonen, Tomi. 2021. *Evolving in Groups: Individualism and Holism in Evolutionary Explanation of Human Social Behaviour*. Helsinki: Helsingin yliopisto.

Krickel, Beate. 2023. "Extended Cognition and the Search for the Mark of Constitution – a Promising Strategy?" In *Situated Cognition Research: Methodological Foundations*, edited by Mark-Oliver Casper and Giuseppe Flavio Artese, 129–146. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-39744-8_8.

McHugh, B., N. Dawson, A. Scrafton, and E. Asen. 2010. "'Hearts on Their sleeves': The Use of Systemic Biofeedback in School Settings." *Journal of Family Therapy* 32 (1): 58–72. https://doi.org/10.1111/j.1467-6427.2009.00486.x.

McKenna, Michael. 2012. *Conversation and Responsibility*. USA: Oxford University Press.

Naeem, H., and J. Hauser. 2024. "Should We Discourage AI Extension? Epistemic Responsibility and AI." *Philosophy & Technology* 37 (91). https://doi.org/10.1007/s13347-024-00774-4.

Neuhäuser, Christian. 2015. "Some Sceptical Remarks Regarding Robot Responsibility and a Way Forward." In *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation*, edited by Catrin Misselhorn, 131–146. Philosophical Studies Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-15515-9_7.

Rupert, Robert. 2009. *Cognitive Systems and the Extended Mind*. New York: Oxford University Press.

Sonne, T., P. Marshall, C. Obel, P. H. Thomsen, and K. Grønbæk. 2016. "An Assistive Technology Design Framework for ADHD." *Proceedings of the 28th Australian Conference on Computer-Human Interaction - OzCHI*, Vol. 16, 60–70. https://doi.org/10.1145/3010915.3010925.

Strawson, Peter F. 1962. "Freedom and Resentment." *Proceedings of the British Academy*, 48:1–25. https://www.thebritishacademy.ac.uk/publishing/proceedings-british-academy/48/strawson/.

Sullins, John P. 2006. "'When is a Robot a Moral Agent?' Edited by Michael Anderson and Susan Leigh Anderson." *The International Review of Information Ethics* 6 (12): 23–30. https://doi.org/10.1017/CBO9780511978036.013.

Telakivi, Pii. 2023. *Extending the Extended Mind: From Cognition to Consciousness*. Cham, Switzerland: Palgrave Macmillan.

Vold, Karina, and José Hernández-Orallo. 2022. "AI Extenders and the Ethics of Mental Health." In *Ethics of Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*, edited by M. Ienca and F Jotterand, 177–202. Cham, Switzerland: Springer.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, Massachusetts: Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674766235.

Wheeler, Michael. 2019. "The Reappearing Tool: Transparency, Smart Technology, and the Extended Mind." *AI & Society* 34 (4): 857–866. https://doi.org/10.1007/s00146-018-0824-x.